Real-Time Face Pose Estimation

Stephen J. McKenna and Shaogang Gong

S. J. McKenna is with the Department of Applied Computing, University of Dundee, Dundee DD1 4HN, Scotland. E-mail:stephen@dcs.qmw.ac.uk. S. Gong is with the Department of Computer Science, Queen Mary and Westfield College, London, England. E-mail: sgg@dcs.qmw.ac.uk

Abstract

Methods were investigated for estimating the poses of human faces undergoing large rotations in depth. Dimensionality reduction using principal components analysis enabled pose changes to be visualised as manifolds in low-dimensional subspaces and provided a useful mechanism for investigating these changes. Appearance-based matching using Gabor wavelets was developed for real-time face tracking and pose estimation. A real-time Gabor wavelet projection was implemented using a Datacube MaxVideo 250 whilst an alternative system for real-time pose estimation used only standard PC hardware.

Keywords

View-based representation, Appearance-based correspondence, Dimensionality reduction, Principal components analysis, Gabor wavelets, Face rotation in depth, Pose eigen-spaces, Real-time pose estimation, Face tracking, Face template matching.

I. Introduction

View-based object representations using sets of 2D views rather than explicit 3D models are becoming increasingly attractive for computer vision. The approach is motivated on psychophysical, neurobiological [1], [2] and computational grounds. Its popularity is partly due to the fact that computation can be made simple by avoiding the need to build 3D models or to perform explicit 3D reconstruction. This “simplicity” can also facilitate the engineering of real-time systems for tracking and analysis of complex 3D objects. In addition, view-based representations do not directly encode prior knowledge of 3D shape. An important consequence is that they can be learned directly from a (possibly labelled) set of images. In this paper, the objects of interest are human faces. Although face detection and recognition have been widely studied, faces are usually constrained to frontal or near-frontal views. A human head rotating in depth (out of the image-plane) induces non-linear transformations in the projected image of the face. Facial features become occluded and the outline of the face alters its shape causing interference with the background. Pose estimation is therefore a difficult task.

This paper is concerned with real-time pose estimation using view-based representation and appearance-based correspondence for matching faces in the presence of such non-linear transformations. A real-time pose estimator can be used to drive graphical models (avatars) for applications such as virtual teleconferencing. It can also be used to index a more detailed view-specific representation for identity recognition and expression analysis. In addition, pose prediction is useful for overcoming display-lags in real-time interactive and visual communication applications.

Two real-time systems are presented in this work. The first uses specialised hardware (a Datacube MaxVideo 250) to implement an approximation to a Gabor wavelet projection (GWP). The second performs face tracking and pose estimation using only a standard PC and a low-end frame grabber. Eye features were used to align views across rotations in depth by exploiting facial symmetry. Appearance-based matching was used to perform tracking and pose estimation.

In the next section, the issue of alignment in a view-based representation is addressed. The problem of face tracking and pose estimation is then defined in more detail in section III where some previously suggested methods are also reviewed. Section IV describes the use of principal components analysis (PCA) to examine face pose distributions. In section V, the GWP is described for appearance-based correspondence and the resulting pose transformations are examined using PCA in section VI. Section VII describes a real-time pose estimator. Specialised hardware was used to implement an approximate GWP. An alternative system used only standard PC hardware to perform face tracking and pose estimation for a virtual teleconferencing application. Conclusions are drawn in section VIII.

II. View Alignment: Feature-based versus Appearance-based Correspondence

Alignment is crucial for a view-based representation. This can be achieved by establishing either feature-based or appearance-based correspondence. In theory, a view-based representation consists of a linear vector space in which each view is represented by a vector [3]. This is only computationally valid if images are "aligned" in the vector space. Ideally, this can be achieved by establishing exact correspondences between the pixels across the set of views [4]. However, dense correspondences are difficult to estimate and rotations in depth result in self-occlusions which prohibit complete sets of image correspondences from being established. Such an approach suffers from similar computational difficulties to the establishment of "morph fields" [5]. In practice, rotation in depth forms non-linear manifolds which can be approximated locally as linear vector spaces [6]. Alternatively, correspondences can be established for only a restricted set of facial features. For example, Gabor wavelets have been used to establish exact correspondences for these facial feature points and this has enabled facial feature tracking [7]. Images can then be "warped" to a canonical shape in order to separate shape (features' image co-ordinates) and texture (photometric properties) into separate linear vector spaces [6]. The shape can be used to recover pose [8]. We refer to this form of alignment as *feature-based.*

An alternative approach uses "anchor" points such as the locations of the two eyes or simply the entire face as a "template", to bring views into alignment by translation, rotation and scaling in the image-plane [9]. This is computationally not only desirable for real-time but also valid. Although correspondences established with such an *appearance-based* approach are only approximate (since they are not point-wise), the vector space formed by faces at a specific view is of low dimensionality [10], [11]. In other words, it is reasonable to assume that it is not necessarily always the case that dense correspondence has to be established in order to obtain the necessary alignment for view-based representation.

The methods described here do not attempt to establish exact correspondences. Instead, an appearance-based approach has been adopted in which spatial filtering is used to construct templates and to compensate for inexact correspondences. This filtering is based upon a GWP and has several additional advantages. It is used to investigate the role of locally oriented features at a range of spatial frequencies in selecting face pose. It also provides some invariance under changes in illumination conditions, skin tone and hair colour. PCA was used to visualise the manifold described by rotation in depth and to investigate the use of a GWP face representation for distinguishing poses. The use of PCA here was analogous to the parametric eigenspaces of Murase and Nayar [12].

III. Face Tracking and Pose Estimation

Face tracking and pose estimation entail the recovery from each face image of a 6 dimensional parametric vector $\mathbf{P} = (x, y, s, r_x, r_y, r_z)$, where (x, y) is the image-plane position of the face, s is its scale or projected size, (r_x, r_y) is the head's rotation in depth and r_z is the rotation of the head in the image-plane. A "nodding" head undergoes x-axis rotation whilst a "shaking" head undergoes y-axis rotation. The image transformations induced by changing the values of the parameters x, y, s and r_z can all be approximated using linear (affine) image-plane transformations. However, as already noted, rotations in depth result in non-linear transformations.

Face detection is needed in order to bootstrap the tracking and pose estimation process. Methods for face detection usually assume frontal or near-frontal views and tend to be computationally expensive. They have been based upon colour [13], silhouette, spatial configuration of facial features and pattern recognition techniques (see [14] for references). Face detection is not the main topic of this paper and the systems described here used an appearance-based matching scheme for detection and tracking. Further references on methods for face processing can be found in reviews [15], [16].

A. Feature-based Correspondence

Matching methods which rely upon maintaining correspondences between facial features inevitably encounter problems due to self-occlusion, background interference and unreliable feature tracking. They also require high-resolution images of the face.

Two types of feature-based matching are possible. The first kind relies upon the ability to locate and track image features known to correspond to features on the face itself [17], [18], [19], [20]. Self-occlusion restricts such methods to local regions of the view-sphere. The second type also relies upon locating and tracking suitable features but these features are chosen based only upon intensity characteristics and have no known conceptual meaning [21], [22], [23], [24]. Head pose is usually estimated relative to the pose in the first frame of a sequence which is typically a frontal view.

Gee and Cipolla used geometric face models for pose estimation [17], [18]. These methods relied critically upon accurately detecting the correct facial features and it is not clear how this was done. Kruger *et al.* used computationally intensive elastic graph matching to locate and estimate the pose of faces [19]. The graphs consisted of connected nodes of Gabor filter "jets". Different graph models were needed for different poses, leading to a poorly integrated and computationally expensive approach. Taylor *et al.* used "active shape models" to locate faces and to model their variations [20]. If these statistical models are trained using suitable data, modes of variation can be extracted which correspond to rotations in depth.

Azarbayejani *et al.* used an extended Kalman filter (EKF) to recover head pose from 10-20 tracked feature points [21], [22]. A similar but computationally intensive approach used an ellipsoidal model and a dense optical flow field [23]. Maurer *et al.* tracked feature points using Gabor wavelets and phase-based displacement estimation [24]. In order to estimate head pose, the tracked points were assumed to lie in a plane. It took several seconds per frame to perform the wavelet convolutions. For similar purposes, phase-based displacement estimation using a real-time GWP implementation (see section VII) can be combined with a shape model in order to perform facial feature point tracking [7].

B. Appearance-based Correspondence

Although appearance-based matching has been used for both face recognition [25] and expression analysis [26] of frontal facial views, its use with rotation in depth is of more interest here. In large, there seems to be little reported work on pose estimation using appearance-based techniques. Beymer used masked face templates at 15 different viewing angles to perform identity recognition [27]. Niyogi and Freeman used hierarchical indexing to match many intensity image templates (40 × 30 pixels) of entire heads at different viewing angles [28]. Poses used were $r_x = \{-30^\circ, 0^\circ, 30^\circ\}$ and $r_y = \{-40^\circ, -20^\circ, 0^\circ, 20^\circ, 40^\circ\}$. Nearest-neighbour matching against a uniform background resulted in the correct pose in 48% of the images tested.

IV. Face Pose Eigenspace

In this section, a method for visualising the effects of face rotations in depth is described. In the following sections, this technique is first used to investigate the role of GWP face representations. This in turn motivates the design of the real-time systems for pose estimation.

Sequences of heads rotating from profile to profile under different lighting conditions were obtained as output from a head tracking system described elsewhere [14], [29]. These were 60 frames long and were automatically normalised with respect to translation and scale by the tracker. Each image was also normalised by subtracting its mean intensity and dividing by its standard deviation. This corrected variations in overall illumination intensity, camera gain and imaging aperture. (Factors such as skin tone and hair colour also influence the first and second moments of intensity so this correction is only approximate). An example can be seen in Fig. 1.

Given an n-frame sequence of a head rotating in depth, a pose eigen-space (PES) can be calculated using PCA as follows. Each image frame is of size $m = p \times q$ pixels and defines an m-dimensional column vector $\mathbf{x}$. The sequence defines a set of such vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Furthermore, the mean, $\boldsymbol{\mu}$, and the covariance matrix, $\boldsymbol{\Sigma}$, of this image set are given by:

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \tag{1}$$

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} [\mathbf{x}_i - \boldsymbol{\mu}][\mathbf{x}_i - \boldsymbol{\mu}]^{\mathrm{T}} \tag{2}$$

where $\boldsymbol{\Sigma}$ is an $m \times m$ matrix. Let $\mathbf{u}_j$, $j = 1 \ldots n'$, be the n' eigenvectors of $\boldsymbol{\Sigma}$ which have the largest corresponding eigenvalues λ_j:

$$\boldsymbol{\Sigma}\mathbf{u}_j = \lambda_j \mathbf{u}_j \tag{3}$$

The n' eigenvectors are the principal components and form the axes of a PES. In practice, the covariance matrix $\boldsymbol{\Sigma}$ is singular since $n \ll m$. However, the $n' < n$ eigenvectors can be estimated

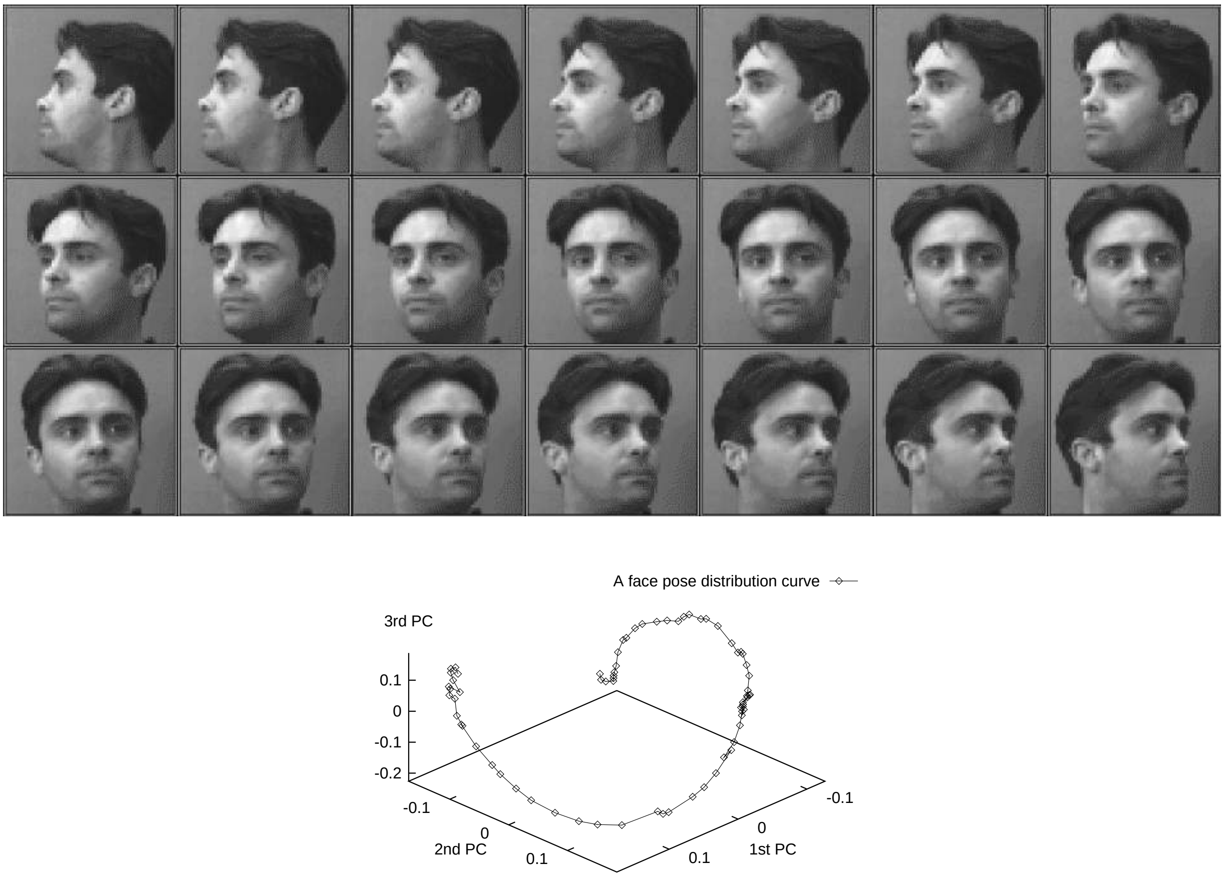

Fig. 1. The pose eigen-space (PES) of a 60-frame sequence of a head rotating from profile to profile. (Every 3rd frame is shown.)

using Singular Value Decomposition [30].

For an image, $\mathbf{x}$, an n'-dimensional "pattern vector", $\boldsymbol{\Omega}(\mathbf{x}) = [\omega_1\, \omega_2\, \ldots\, \omega_{n'}]^{\mathrm{T}}$, can be computed by projection onto each of the eigenvectors $\mathbf{u}_j$:

$$\omega_j = \mathbf{u}_j^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}) \qquad j = 1, \ldots, n' \tag{4}$$

This pattern vector can be normalised by the eigenvalues in order to give the data equal variance along each principal component axis:

$$\boldsymbol{\Omega}_{\mathbf{norm}}(\mathbf{x}) = [\frac{\omega_1}{\lambda_1}\, \frac{\omega_2}{\lambda_2}\, \ldots\, \frac{\omega_{n'}}{\lambda_{n'}}]^{\mathrm{T}} \tag{5}$$

A face sequence can thus be approximated by its pattern vectors and the first n' eigenvectors and eigenvalues. When $n' \leq 3$ the pattern vectors can be plotted on a graph so that the distribution of poses in the representation space can be visualised. Fig. 1 shows an example sequence along with a plot of its normalised 3D pattern vectors, $\mathbf{\Omega_{norm}}$. The pose of a novel face image of the person can be estimated by projecting it into this PES. In the case of unnormalised pattern vectors, Euclidean distance in the PES is a least-squares approximation to Euclidean distance in the image space. The commonly used methods of minimising the sum-of-squared-difference or maximising the correlation between images can therefore be efficiently approximated by minimising Euclidean distance in the PES [31]. If normalised pattern vectors are used, Euclidean distance in PES is related to the Mahalanobis distance.

The following conclusions regarding pose distributions can be drawn from this PES analysis. Projected normalised pattern vectors of head sequences under different lighting conditions form smooth manifolds in a 3D PES. Fig. 2 shows three curves which form a fairly smooth manifold parameterised by pose and illumination. In particular, the 3rd principal component seems to capture changes caused by lighting conditions. This manifold is similar to those obtained by Murase and Nayar who derived parametric eigenspaces from various non-face objects under robotically manipulated pose and illumination conditions [12]. In contrast, the face sequences used here were produced by an automatic visual tracking system with left, right and ambient lighting. As a result, the manifold shown here is less smooth, reflecting more realistic conditions.

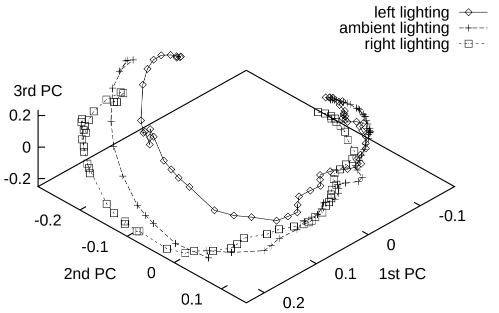

Fig. 2. A manifold formed by face sequences under different lighting conditions rotating from profile-to-profile (−90° to +90°). The manifold is formed by three tracked head sequences of the same person under different lighting conditions. The images were projected into a PES derived from only one of these sequences.

Even after intensity normalisation, variations in illumination conditions are apparent. In the next section, the use of Gabor wavelet projections is described for face representation. This helps to alleviate lighting effects and has several other benefits.

V. GWP Face Representation

A Gabor wavelet projection (GWP) yields images which are locally normalised in intensity and decomposed in terms of spatial frequency and orientation. It thus provides a mechanism for obtaining (a) some invariance under intensity transformations due to illumination, skin tone and hair colour, (b) selectivity in scale by providing a pyramid representation, and more importantly for our studies, (c) it permits investigation into the role of locally oriented features with regard to pose changes.

In a scheme proposed by Würtz [32], a Gabor wavelet transform (GWT) can be performed by convolutions with Gabor kernels in the Fourier domain. A single Gabor function (the mother wavelet) is parameterised by a vector $\mathbf{k}=\binom{k_1}{k_2}$, defining variations in spatial frequency and orientation. Then a GWT in $[-\pi < \boldsymbol{\omega} = \binom{u}{v} < \pi]$ is given by:

$$\mathbf{F_k}(\boldsymbol{\omega}) = \exp\left(-\frac{\sigma^2(\boldsymbol{\omega}-\mathbf{k})^2}{2\mathbf{k}^2}\right) - \exp\left(-\frac{\sigma^2(\boldsymbol{\omega}^2+\mathbf{k}^2)}{2\mathbf{k}^2}\right) \tag{6}$$

The second term in Equation (6) results in "admissibility" i.e. removal of the DC component. A consequence of this is zero response to spatially constant intensity.

However, for computational efficiency using convolution hardware, approximated spatial convolution is more desirable. An approximate Gabor wavelet projection (GWP) of an image was obtained by convolution with a set of 2D Gabor kernels, i.e. sinusoidally modulated 2D Gaussian functions of different spatial frequencies and orientations [33]. Fig. 3 shows Gabor wavelet kernels in the spatial domain at three frequencies and four orientations varying by 45° from 0° to 135°. The kernels, which approximate Equation (6), are defined as follows:

$$K_{odd}(x,y) = \xi \sin\theta e^{-r^2(\frac{k}{\sigma})^2} \tag{7}$$

$$K_{even}(x,y) = \xi(\cos\theta - e^{\frac{-\sigma^2}{2}})e^{-r^2(\frac{k}{\sigma})^2} \tag{8}$$

$$\theta = [kx, k\sqrt{2}(x+y), ky, k\sqrt{2}(y-x)] \tag{9}$$

where $r^2 = x^2 + y^2$, σ controls the width of the Gaussian envelope and k controls the spatial frequency. The extra Gaussian term in K_{even} makes the kernel admissible.

Fig. 4 shows Gabor wavelets parameterised by 3 spatial frequencies and 4 orientations (0°, 45°, 90°, 135°) applied to a face image to yield a GWP. The real and imaginary parts of the

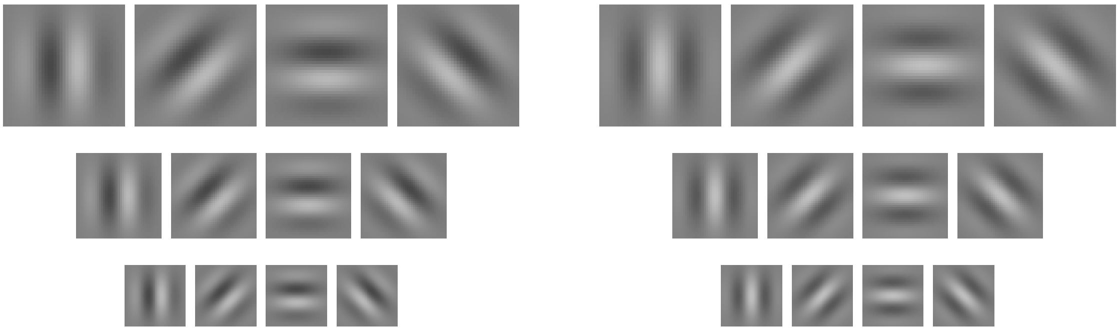

Fig. 3. Gabor wavelet kernels at four orientations and three spatial frequencies. The imaginary (odd) components appear on the left and the real (even) components on the right.

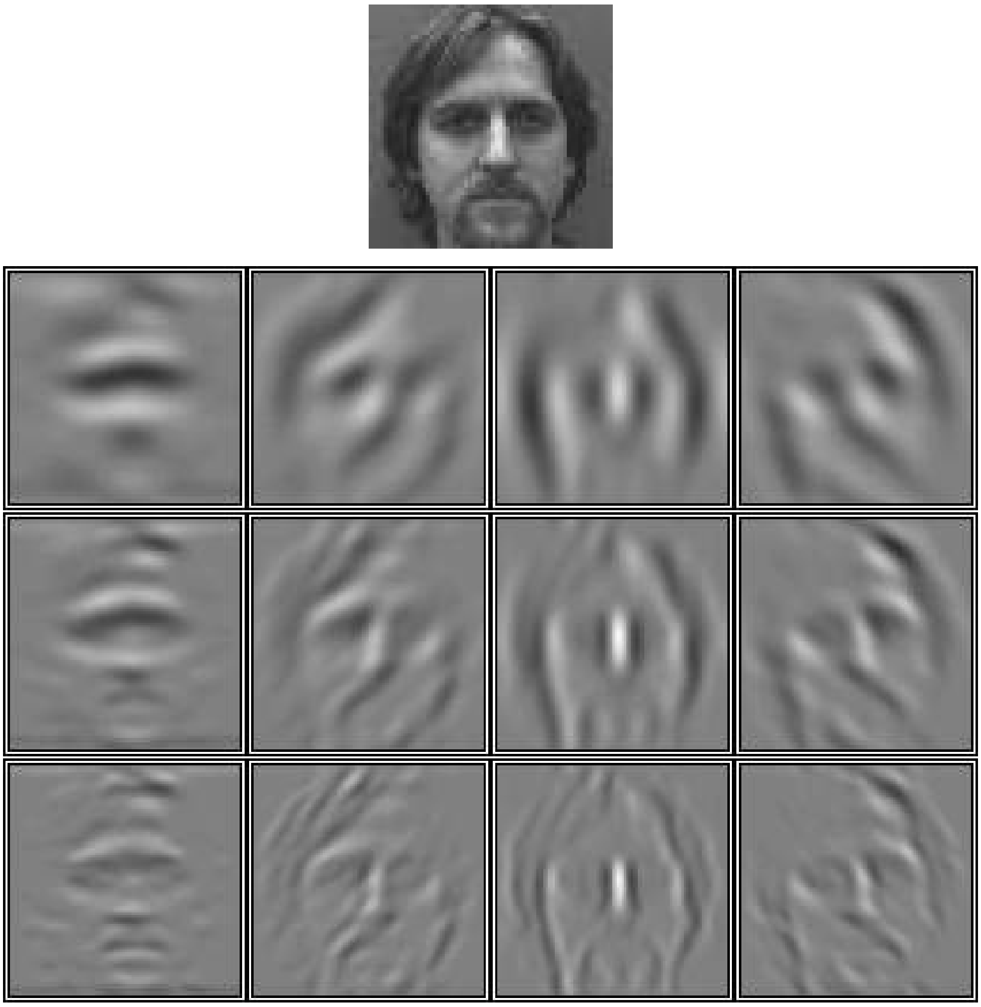

Fig. 4. A face image with GWP responses at three spatial frequencies and four orientations.

kernel responses oscillate with their characteristic frequency making them highly sensitive to image-plane translations and therefore ill-suited to matching. This undesirable property can be avoided by taking the magnitude of the responses thereby removing phase information [34]. Fig. 5 shows the magnitude responses of the GWP to the face image of Fig. 4. Gabor magnitude and phase images, G_{mag} and G_{pha}, are obtained from the even and odd convolution responses, G_{even} and G_{odd}, as follows.

$$G_{odd}(x, y) = I(x, y) * K_{odd} \tag{10}$$

$$G_{even}(x, y) = I(x, y) * K_{even} \tag{11}$$

$$G_{mag}(x, y) = \sqrt{(G_{odd}(x, y))^2 + (G_{even}(x, y))^2} \tag{12}$$

$$G_{pha}(x, y) = \tan^{-1} \frac{G_{odd}(x, y)}{G_{even}(x, y)} \tag{13}$$

The Gabor phase responses rotate approximately with the spatial frequency making them highly predictable and well suited to estimating feature displacement. Such a mechanism was combined with a point distribution model of deformable shape to obtain robust facial feature tracking [7]. However, in an appearance-based approach, inexact correspondences can be compensated by ignoring the phase and considering only the magnitudes of the responses which tend to vary smoothly over an image. The magnitude responses are used in the work described here. At lower frequencies, faces were smoothed to a larger extent resulting in less sensitivity to small translations in the image-plane and greater correlation between nearby images in a sequence. However, using excessively low frequencies results in loss of relevant spatial structure (see the first row in Fig. 5).

A GWP face $G(\mathbf{x})$ is obtained by superimposing the GWP responses. The result is similar to the original intensity image except that intensity distributions are locally normalised. (A "GWP face"-like representation could also be obtained by using symmetric filters.) Furthermore, a "composite" GWP face $CG(\mathbf{x})$ of equal dimensionality to $G(\mathbf{x})$ is formed by concatenating four "oriented" 1/4 sized images, each a sub-sampled (by a factor of four) Gabor response to a different orientation (see Fig. 6). In the next section, we describe the purpose and show the effect of forming such a composite GWP face representation in understanding face pose distributions in a PES.

VI. Pose Distribution of Faces Rotating in Depth

The formation of a PES from a sequence of face images was described in Section IV. If the images used are composite GWP faces, each eigenvector derived from this representation can be

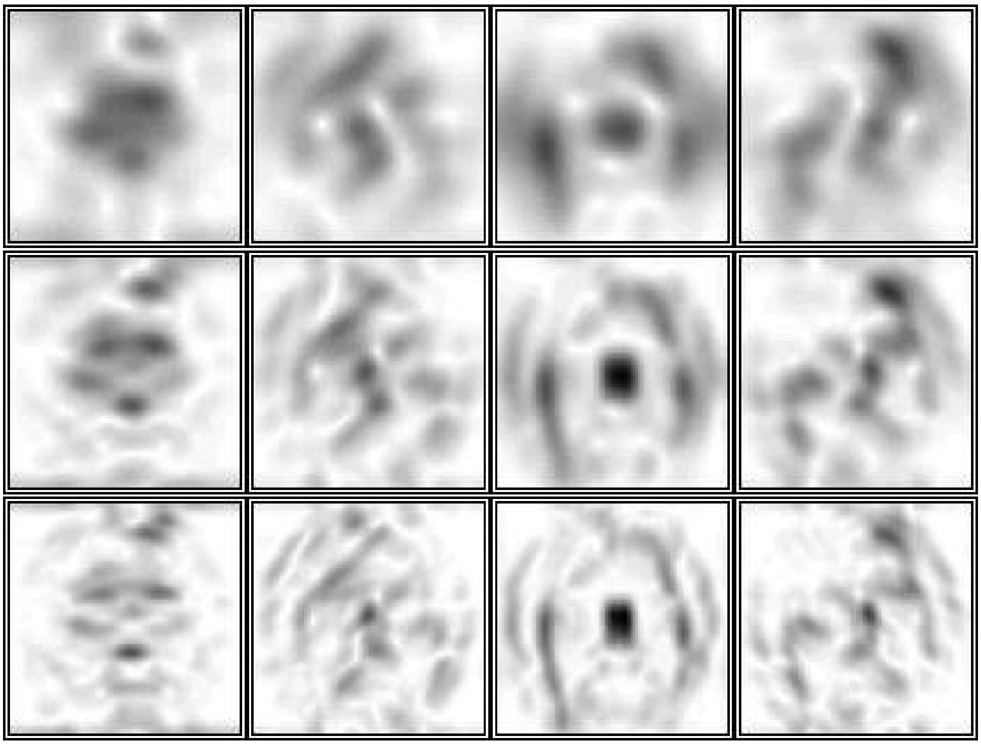

Fig. 5. GWP magnitude responses of the face images in Fig. 4.

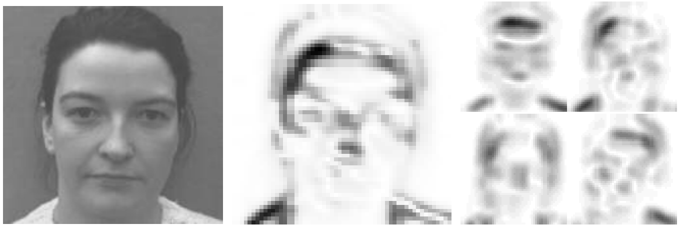

Fig. 6. Left: a normalised intensity face I. Centre: a GWP face $G(\mathbf{x})$. Right: a composite GWP face $CG(\mathbf{x})$ of equal dimensionality.

visualised as a composite eigen-image consisting of four oriented sub-images. The magnitude of each pixel in such an eigen-image can be envisaged as a measure of the variability of the response of one complex Gabor kernel centred at the corresponding position in the original image. In particular, the magnitudes of the first eigen-image would indicate *where* in the image-plane *which* orientations encode the most information about pose. This gives us a means to investigate the role of locally oriented features in pose changes.

Here, PES manifolds are derived from GWP-face and composite GWP face images. These are visually compared to those obtained using intensity image sequences in a 3D PES. Only a single spatial frequency was used for the GWP in order to simplify the computation.

A. Data Preparation

A set of pose-labelled face sequences of 12 people were obtained under controlled conditions in which subjects were asked to look at markers on the wall positioned at angles from 0° (frontal view) to 90° (right profile view) in 10° increments. Profile-to-profile sequences were generated by mirroring the sequences so that each sequence consisted of 19 frames of known poses. Illumination varied between sequences. All images were sub-sampled with smoothing to 64×64 pixels. Fig. 7 shows 5 frames from such a sequence.

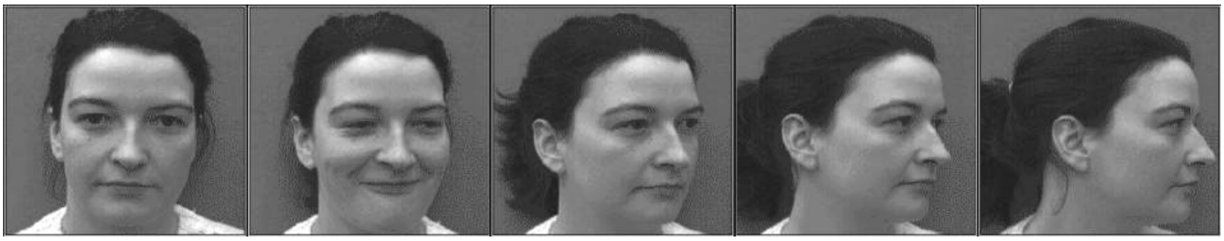

Fig. 7. A head rotates in depth.

In order to measure the effects of pose change, other degrees of freedom such as image-plane translations and scale changes are carefully removed by manually cropping the images. An important point to note is that rotation of a head results in a horizontal translation of the face in the image-plane. This makes the alignment of images of different poses rather difficult. Initially, the approach of alignment by facial features was adopted. This can result in a sequence in which the "centroid" of the head translates horizontally as the head rotates in depth. Alignment based on establishing correspondences as discussed in Section II becomes problematic due to occlusions. For all the experiments described in this section, images were aligned approximately around the visual centroid of the head, automatically for the tracked sequences and manually for the labelled sequences.

B. PES of Intensity Faces

A generic PES was derived using a mean sequence, $\boldsymbol{\mu} = \{\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_{n-1}\}$, formed by taking the mean normalised intensity image at each of n pose angles over many different face sequences. The plot in Fig. 8 shows the pose distribution of a mean sequence formed using 11 face sequences of different people. Also plotted are the projections into this mean PES of a face sequence of a novel person and a non-face sequence of a fan rotating similarly from profile-to-profile. Since the faces were not perfectly symmetric, mirroring has caused a discontinuity at the frontal view.

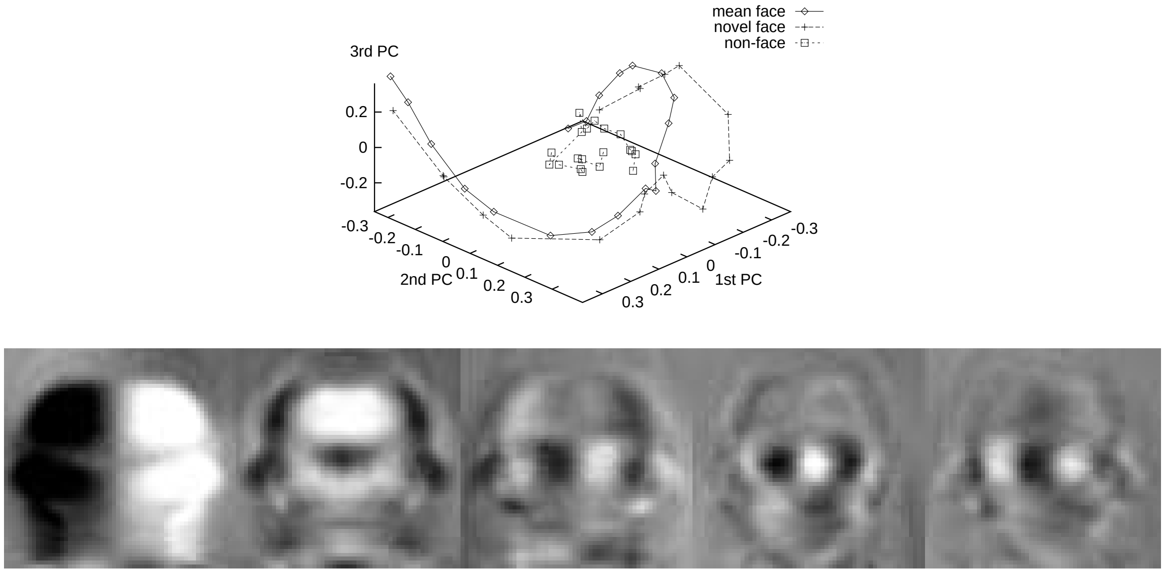

Fig. 8. (1) Plot: A 3D PES was formed from the mean face sequence. Pattern vectors were plotted for the mean face sequence, a novel face sequence and a non-face object (a fan) rotating from −90° to +90°. (2) Bottom row: the first 5 PC's of the mean face sequence.

The following observations were made regarding this mean intensity PES. Whilst the 1st principal component (PC) separates the left and right poses, the 2nd and 3rd PCs jointly discriminate between poses from profile to frontal views reasonably well. This can also be observed from the eigen-images shown beneath the plot. It is worth noticing that although higher order PCs have not been plotted in this 3D PES, it is clear that the 4th and 5th PCs capture finer changes in pose angle. In principle, the poses of all the frames of the novel face sequence can be computed by finding the nearest point along the mean curve. This is an efficient approximation to minimising sum-of-squared-difference or maximising correlation between a novel face and a mean face of known pose. For the same reason, the non-face object is distant from the face curves in the PES for most poses.

C. PES of GWP Faces

In order to examine the effect of using GWP faces (see the second image in Fig. 6) on pose distribution, a PES based on a mean GWP face sequence was derived. Similarly to the mean intensity PES, a mean GWP face sequence, $\boldsymbol{\mu}_g = \{\bar{\mathbf{g}}_0, \bar{\mathbf{g}}_1, \ldots, \bar{\mathbf{g}}_{n-1}\}$, was obtained by taking the mean GWP face at each pose angle over 11 sequences of different people. Fig. 9 shows the pose distribution curve of this mean GWP sequence and the projections of two different GWP face sequences into this PES. One important observation is that compared with the mean intensity PES shown in Fig. 8, the pose distributions in both 2nd and 3rd PC dimensions are more linear. This may be due to the fact that the GWP faces are less sensitive to changes in illumination and differences in local features. However, PES of GWP faces is more sensitive to translations in the image-plane.

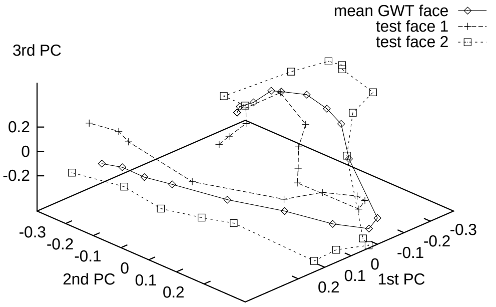

Fig. 9. Pose distribution curves of (1) the mean GWP face representation of 11 face sequences (2) two test GWP face sequences. All 3 are projected into the mean GWP PES.

Although it is clear from those eigen-images shown in Fig. 8 that the 1st PC plays an important role in dividing the pose sphere into two groups and that the subsequent PCs encode higher frequency changes caused by pose variation, it is not clear what effect local oriented facial features have on the pose distribution. In order to examine this issue, a mean composite GWP sequence, $\boldsymbol{\mu}_{cg}$ was constructed similarly to $\boldsymbol{\mu}_g$. Fig. 10 shows the pose distribution curves in the PES of this mean composite GWP face sequence. Compared to both the PES of the mean intensity (Fig. 8) and the PES of GWP faces (Fig. 9), the pose distribution curves are well linearised. The pose angles are symmetrically distributed along two lines, clearly separable and much easier to measure. The discontinuity at the frontal view is due to the fact the sequences were formed by mirroring images (see Section VI-A).

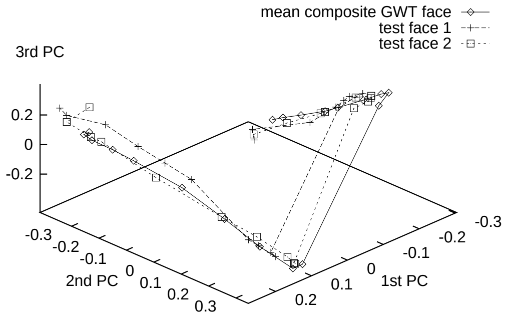

Fig. 10. Projections of the mean composite GWP face sequence and two test face sequences into the mean PES.

VII. A Real-Time Pose Estimator

A real-time system for head pose estimation was developed based upon the findings on pose distribution. The scenario considered is that of a user sitting in front of a monitor. The background is cluttered and may contain moving people and objects. The allowed range of head rotations in depth should be at least those poses in which the user can see the monitor. Head pose needs to be estimated to a precision which is perceptually acceptable to users envisaged in a virtual teleconferencing environment. Users will usually already have been required to "log in" to the computer system and the head tracker can take advantage of this by loading separate models for each user. Under these conditions, a person-specific, appearance-based model has been developed to obtain real-time performance with modest hardware (a 133MHz Pentium PC). The output parameters $\mathbf{P} = (x, y, s, r_x, r_y, r_z)$ from the model have been used to drive an avatar for use in a virtual teleconferencing application [9].

A. Real-time GWP

Firstly, an approximation to a GWP was implemented using specialised pipeline hardware (a Datacube MaxVideo250) in order to achieve real-time performance. Fig. 11 illustrates this implementation. Two real-time 8×8 convolver units and 8 convolution kernels were used: both even and odd kernels at 4 orientations. All kernels were designed to be admissible (see Section V). The even and odd responses were converted to magnitude and phase responses in hardware using a look-up table (LUT). Different spatial frequencies were obtained by sub-sampling the image prior to convolution. In this implementation, suitable parameters were found to be $\sigma = 2.0$,

$\xi = 64$ and $k = \frac{5}{16}\pi$ (see Equations (7) and (8)).

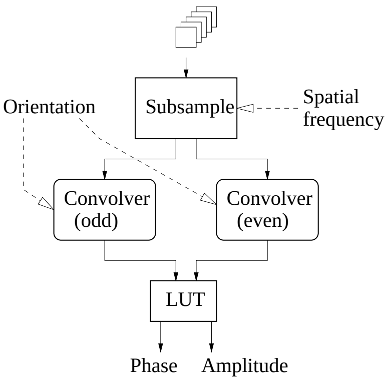

Fig. 11. Datacube implementation of GWP.

Secondly, an alternative system which performed all computation including the convolutions in software was implemented on a 133MHz Pentium PC. In this case, real-time performance was achieved by using 3×3 convolution kernels to obtain simple oriented responses. Monochrome video input from a Matrox Meteor frame grabber was sub-sampled after Gaussian smoothing to a resolution of 96×72 pixels. Performance without smoothing was unacceptable. System bootstrapping and face tracking were performed using templates which were typically 25×30 pixels. The use of such low resolution was motivated not solely by the need for real-time performance. If adequate pose estimation is achievable at this resolution it would seem misguided to require the processing of additional high-frequency information. The human visual system seems able to satisfactorily estimate head pose at such low resolutions.

B. Face Tracking using Template Matching

Model face templates were designed so as to exclude the background to as great an extent as possible without the need for specially shaped masks for each pose (*cf.* [27], [35]). The templates contained most of the visible interior facial region. The hair was largely excluded in order to avoid difficulties with unpredictable and changing hair-styles. Templates were of a fixed size in order to facilitate the use of image interpolation techniques. Such techniques also require the templates for the different poses to be spatially aligned both for reasons stated in Section II and in order to obtain a smooth manifold in image space. This alignment was achieved using the eyes as "anchor" points.

A set of view-based model templates was obtained off-line. Firstly, a frontal view ($r_x = r_y = 0$)

was captured interactively by specifying with a mouse the points on each eye nearest to the side of the head: (x_{leye}, y_{leye}) and (x_{reye}, y_{reye}). The width of all templates was set to $w = x_{reye} - x_{leye}$ and the height to $h = 1.2w$. A rectangular face template was captured with upper-left vertex at $(x_{leye}, [y_{reye} - y_{leye}]/2 + 0.2h)$. Subsequent templates were captured by specifying with a mouse the most extreme point on the eye nearest to occlusion by the head, (x_{occ}, y_{occ}), and then capturing a template with an upper vertex at $(x_{occ}, y_{occ} + 0.2h)$.

To bootstrap the system, a head was detected and tracked using normalised cross-correlation with the view templates. Efficient hierarchical matching was performed using an image pyramid as implemented in the Matrox Imaging Library [36]. The matching score r between an image patch $\mathbf{x}$ and a template model $\mathbf{m}$ is given by:

$$r = \frac{\mu_{\mathbf{xm}} - \mu_{\mathbf{x}}\,\mu_{\mathbf{m}}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{m}}} \tag{14}$$

The tracker was initially trained using raw intensity images and as expected performance was found to degrade drastically after periods typically of minutes due to changes in illumination since the camera was situated near an exterior window. However, the same system using filtered images did not noticeably degrade over a period of several weeks. It is worth pointing out that horizontally oriented kernels respond strongly to the mouth, nostril area, eyes and eyebrows. The responses are most sensitive to x-axis head rotation and images filtered with these kernels are therefore suited to estimating r_x. Since these filters respond to the main facial features, they are also useful for detecting and tracking the face. Vertically oriented kernels respond strongly to the sides of the head and nose. Their response is sensitive to y-axis rotation making them useful for estimating r_y.

C. Temporal Continuity in Pose Estimation

Temporal continuity was exploited in order to restrict the space in which to search for the best template match. In any given frame, the search was centred on the parameters $\mathbf{P} = (x, y, s, r_x, r_y, r_z)$ predicted for that frame. $\mathbf{P}$ was defined in Section III. In the implementation described here, scale and image-plane rotation were assumed to be approximately constant ($s = C$, $r_z = 0$). The extent of the search space around $\mathbf{P}$ was initialised to $\mathbf{S_0} = (d_x, d_y, 0, \theta_x, \theta_y, 0)$. If no strong match was found in a frame, the search space was expanded: $\mathbf{S} = \mathbf{S} + (\Delta d_x, \Delta d_y, 0, \Delta\theta_x, \Delta\theta_y, 0)$. Whenever a strong match was found the search space was reinitialised: $\mathbf{S} = \mathbf{S_0}$. Suitable search parameters for the current implementation were found to be $d_x = 9$, $d_y = 6$, $\Delta d_x = 2$, $\Delta d_y = 1$ and $\theta_x = \theta_y = \Delta\theta_x = \Delta\theta_y = 20°$.

Only a subset of the templates was typically used in any given frame thereby increasing the frame-rate and improving robustness. In general, increasing the number of templates used made a false match more likely. However, decreasing the number of templates made finding no good match a more probable outcome. The search parameters θ_x and θ_y were set to address this trade-off.

A further trade-off exists in setting d_x and d_y. Large values slow down the achievable frame-rate which in turn increases the visual motion between frames. Small values allow faster frame-rates but might not allow the head to be found. In the simplest version of the system, the predicted state $\mathbf{P}$ for frame t was just the estimated state for frame $t - 1$. This approach permitted successful tracking.

The pose estimates obtained using nearest-neighbour template matching were imprecise due to the coarse quantization of the view-sphere used. In addition, the nearest-neighbour template was not always the template with the pose nearest to the current face pose. The pose estimates might be improved by exploiting spatial continuity (see Section VII-G). However, temporal continuity also provides a powerful constraint and can be easily exploited. A simple “moving average” filter on the pose estimates r^* had a significant effect on their perceptual acceptability:

$$r^*(t) = (1 - \alpha)\, r(t) + \alpha\, r^*(t - 1) \tag{15}$$

where α is a constant (typically $\alpha = 0.5$) and $r(t)$ is the pose measured at time t.

D. Tracking r_y

Vertically filtered templates proved useful for estimating y-axis rotation. A suitable interval between templates was found to be 15° for such templates. It should be noted that the intervals on the view-sphere given here are approximate. It is difficult to accurately label the head pose without the use of a calibration method such as a Polhemus tracker. Subjects can be asked to point their heads towards labelled points on the wall but such data labelling is always noisy. Fig. 12 shows a set of 7 templates in the range ±45° which were used to estimate r_y. Fig. 13 shows example frames during tracking and pose estimation using these templates. The estimation of r_y is in good agreement with human perception and shows a significant amount of invariance to r_x.

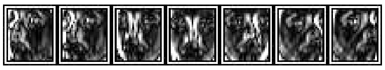

Fig. 12. Templates at 15° intervals used to estimate rotation about the y-axis. The template in the centre is from a frontal view.

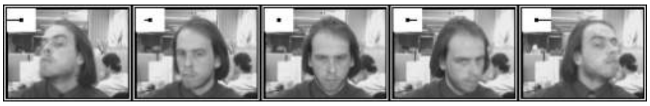

Fig. 13. Tracking and estimating y-axis rotation using the 7 templates of Fig. 12. The needle in the upper-left corner of each image indicates the estimated head rotation.

E. *Tracking r_x*

Horizontally filtered templates proved useful for estimating x-axis rotation. Fig. 14 shows 7 templates at intervals of 10°. Fig. 15 shows example frames from a head being tracked using these templates. The estimates of r_x are in good agreement with human perception and are not adversely affected by small amounts of y-axis rotation. The last frame in Fig. 15 shows the pose estimation breaking down for large r_y. Certain facial expressions caused changes in pose estimates. For example, lowering the eye-brows resulted in too large an estimate of r_x. This expression made the face appear foreshortened as if tilted backwards. Conversely, raising the eyebrows often resulted in too small a value for r_x. Human perception of r_x may be susceptible to similar effects.

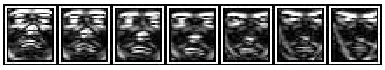

Fig. 14. Templates at 10° intervals used to estimate r_x.

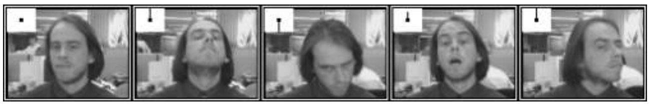

Fig. 15. Tracking and estimating x-axis rotation using the 7 templates of Fig. 14. The last image shows an example of the pose estimation breaking down under large y-axis rotation.

F. *Simultaneous Tracking of (r_x, r_y)*

Horizontally filtered templates were found to be suitable for tracking r_x and r_y simultaneously. Typically, 11 or 15 templates were used (see Fig. 16). Temporal filtering of the pose estimates was needed to overcome the coarse quantization of the view-sphere and the occasional incorrect match. Some example frames showing tracking using this technique are shown in Figs. 17 and 18.

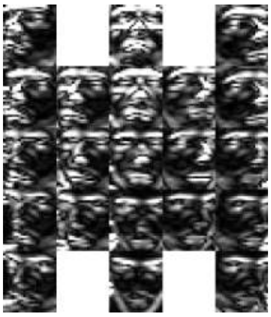

Fig. 16. A set of 21 horizontally filtered templates.

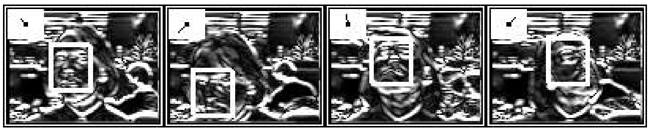

Fig. 17. A head is tracked using templates filtered with a horizontally oriented kernel. The best matching templates' bounding boxes are shown overlaid on a filtered sequence.

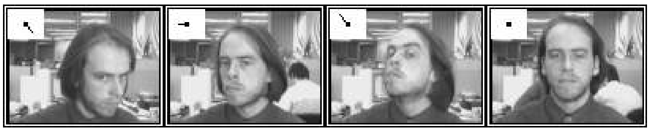

Fig. 18. An unfiltered sequence is shown here for visualisation. The pin diagram indicates estimated pose.

In order to track both r_x and r_y with good precision, the vertically and horizontally filtered images can be combined to provide sensitivity to rotations around both the x and y axes. Two ways in which to combine them are:

1. Both vertical and horizontal responses are used to perform matching and pose estimation in each frame.

2. The vertical and horizontal responses are "interlaced" in time. Matching is performed using vertical responses in the odd numbered frames and using the horizontal responses in the even numbered frames.

The second approach was adopted for the PC platform. This combined the benefits of horizontal responses for face detection and x-axis rotation with those of vertical responses for y-axis rotation. There was no loss of frame-rate. In the case of the Datacube implementation of a GWP with multiple orientations, robustness can be improved by performing template matching with responses at all orientations.

G. RBF networks for pose estimation

The template matching described so far measures the head pose (r_x, r_y) in a given frame using nearest neighbour matching, i.e. the pose label associated with the best matching template is used as the measurement. A more general method is to interpolate over the matching scores of a set of templates. Radial basis function (RBF) networks provide one possible approach to achieving this interpolation [37], [38].

Several RBF networks were trained and their ability to interpolate between templates of different head poses was investigated. Once a match was found for a view template in the current image, an RBF network was applied to the matching image patch. Each hidden unit measured a Gaussian weighted distance to a view template. The output layer was linear and its weights were set using Singular Value Decomposition [38]. Network training was very fast (a few seconds). However, no improvement could be demonstrated over the nearest neighbour "winner takes all" method of pose estimation.

H. Driving an avatar

The pose estimates were used to drive a synthetic head model which rotated in depth around its x and y axes. The result was perceptually acceptable to the user, especially when pose prediction was used to compensate for the time-delay between movement of the real head and the synthetic head [9].

VIII. Conclusions

In this paper, the issue of measuring face pose of a moving head in real-time has been addressed. A composite face representation scheme based on a Gabor wavelet projection (GWP) was introduced in order to both normalise intensity and scale and to investigate the role of locally oriented features in regularising pose distributions. Pose eigenspaces based on principal components analysis were used to represent and interpret the distribution of pose changes from continuous face sequences of rotations in depth. In particular, it was shown that pose changes of a continuous face rotation in depth form a smooth curve in pose eigenspace. Whilst the first principal component (PC) of this eigenspace divides all poses from profile-to-profile into two symmetric parts centred at the frontal view, the remaining PCs differentiate poses between profile to frontal views. The third PC also seems to capture changes in illumination. Furthermore, it seems that the pose distribution curves of faces in the pose eigenspace are distinctively different from those of non-face objects. Although GWP representation reduces the complexity of pose distributions, it is sensitive to translational changes in the image-plane. More interestingly though, the composite GWP representation gives a highly linear pose distribution. It appears that the Gabor kernels of different orientation play some role in "regularising" pose distributions. This is computationally attractive for determining poses of novel faces. Based on these findings, an appearance-based and computationally efficient model for tracking and estimating the pose of a moving human head has been developed. Oriented spatial filters were used to obtain robustness under changing illumination conditions and to improve pose estimation by extracting directionally sensitive facial features. A set of face templates was used to sample the view-sphere. Exploitation of temporal constraints provided robust tracking and pose estimation with an accuracy which was shown to be perceptually acceptable for applications in virtual teleconferencing.

There are several obvious ways in which the system could be extended in the future. Firstly, it could be modified to cope with changes in scale. The area of the image to be searched would be re-sampled at different scales prior to applying the oriented spatial filters. The same set of view-templates could then be used to search at a range of scales. This scale range would also be restricted using temporal continuity.

Secondly, given the estimated poses of continuous head movement, analysis of the facial region at a higher resolution becomes feasible and estimation of non-rigid deformations such as facial expression and mouth shape can be performed with fewer constraints. Most current methods for estimating these non-rigid deformations are limited to frontal or near-frontal views, e.g. [39]. The availability of rigid head pose should facilitate their extension to a wider range of poses with graceful degradation rather than complete failure under large rotations in depth.

Thirdly, the system currently requires one to train one's own head model interactively before the system starts to function. This process involves capturing a few templates by rotating the head and "clicking" on an eye feature. It takes about 2 minutes. It was not possible to explore a person independent model due to the lack of any suitable database with labelled head pose angles. In order to derive a model which reliably exhibits invariance to identity, a representative database of different people at various poses would be required. If such a database became available, each view-template could be replaced by a statistical model capturing the variations due to identity. The method developed here should be extendable via training-from-examples techniques. Work is currently being undertaken to explore the use of a Polhemus tracker for collecting pose ground-truth.

IX. Acknowledgements

This research was funded by EPSRC Grant No. GR/K44657, EC Grant No. CHRX-CT94-0636 and a BT Short-Term Fellowship. The authors gratefully acknowledge the assistance of J. J. Collins (now of University of Limerick, Ireland) who contributed extensively to many of the experiments described in sections IV-VI. They also wish to thank Rolf Würtz for many helpful discussions. Some of the work was carried out during S. J. McKenna's Short-Term Research Fellowship at BT Labs., Martlesham Heath, England. He wishes to thank Dave Machin and Phil Sheppard for their hospitality and helpful discussions. Finally, thanks are due to all those who volunteered to have their face images captured.

References

[1] H. Bülthoff, S. Edelman, and M. Tarr, "How are three-dimensional objects represented in the brain?," AI Memo 1479, MIT, Cambridge, Massachusetts, April 1994.

[2] N.K. Logothetis, J. Pauls, and T. Poggio, "Spatial reference frames for object recognition: Tuning for rotations in depth," AI Memo 1533, MIT, Cambridge, Massachusetts, March 1995.

[3] S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 992–1006, October 1991.

[4] D. J. Beymer and T. Poggio, "Image representations for visual learning," *Science*, vol. 272, pp. 1905–1909, 28 June 1996.

[5] M. Bichsel, "Automatic interpolation and recognition of face images by morphing," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Killington, Vermont, October 1996, pp. 128–135.

[6] I. Craw, "A manifold model of face and object recognition," in *Cognitive and Computational Aspects of Face Recognition*, T. R. Valentine, Ed., pp. 183–203. Routledge, London, 1995.

[7] S. J. McKenna, S. Gong, R. P. Wurtz, J. Tanner, and D. Banin, "Tracking facial feature points with Gabor wavelets and shape models," in *Int. Conf. on Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science 1206*, Crans-Montana, Switzerland, 1997.

[8] T. Poggio and S. Edelman, "A network that learns to recognize three-dimensional objects," *Nature*, vol. 343, January 1990.

[9] S. J. McKenna, "View-based estimation of head pose," Tech. Rep. IA377656, BT, Advanced Perception, BT Labs., Martlesham Heath, England, 1996.

[10] M. Kirby and L. Sirovich, "Application of the Karhunen-Loéve procedure for the characterization of human faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, 1990.

[11] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, Seattle, July 1994.

[12] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *Int. J. Computer Vision*, vol. 14, 1995.

[13] Y. Raja, S. J. McKenna, and S Gong, "Tracking and segmenting people in varying lighting conditions using colour," in *3rd International Conference on Face and Gesture Recognition*, Nara, Japan, 1998.

[14] S. J. McKenna and S. Gong, "Tracking faces," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, Killington, VT., 1996.

[15] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces, a survey," *Proc. IEEE*, vol. 83, pp. 705–740, 1995.

[16] D. Valentin, H. Abdi, A. J. O'Toole, and G. W. Cottrell, "Connectionist models of face processing - a survey," *Pattern Recognition*, vol. 27, no. 9, pp. 1209–1230, 1994.

[17] A. H. Gee and R. Cipolla, "Determining the gaze of faces in images," *Image and Vision Computing*, vol. 12, no. 10, pp. 639–647, 1994.

[18] A. H. Gee and R. Cipolla, "Fast visual tracking by temporal consensus," *Image and Vision Computing*, vol. 14, no. 2, pp. 105–114, 1996.

[19] N. Kruger, M. Potzsch, T. Maurer, and M. Rinne, "Estimation of face position and pose with labeled graphs," in *British Machine Vision Conference*, Edinburgh, 1996.

[20] A. Lanitis, C. J. Taylor, and T. F. Cootes, "A unified approach to coding and interpreting face images," in *IEEE Int. Conf. Computer Vision*, Cambridge, Mass., 1995, pp. 368–373.

[21] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland, "Visually controlled graphics," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, 1993.

[22] A. Azarbayejani and A. Pentland, "Recursive estimation of motion, structure and focal length," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, 1995.

[23] S. Basu, I. Essa, and A. Pentland, "Motion regularization for model-based head tracking," in *ICPR*, 1996.

[24] T. Maurer and C. von der Malsburg, "Tracking and learning graphs and pose on image sequences of faces," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, 1996, pp. 176–181.

[25] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[26] T. Darrell, I. Essa, and A. Pentland, "Task-specific gesture analysis in real-time using interpolated views," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1236–1242, 1996.

[27] D. J. Beymer, "Face recognition under varying pose," AI Memo 1461, MIT, Cambridge, Massachusetts, 1993.

[28] S. Niyogi and W. T. Freeman, "Example-based head tracking," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, Killington, VT., 1996.

[29] S. J. McKenna, S. Gong, and H. Liddell, "Real-time tracking for an integrated face recognition system," in *Second European Workshop on Parallel Modelling of Neural Operators*, Faro, Portugal, November 1995.

[30] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, Cambridge Press, Cambridge, England, 1992.

[31] S. K. Nayar, H. Murase, and S. A. Nene, "Parametric appearance representations," in *Early Visual Learning*, S. K. Nayar and T. Poggio, Eds., chapter 6. Oxford Press, 1996.

[32] R. P. Würtz, *Multilayer Dynamic Link Networks for Establishing Image Point Correspondences and Visual Object Recognition*, Verlag Harri Deutsch, 1994.

[33] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. of Optical Society of America*, vol. 2, 1985.

[34] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition and gender determination," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Zurich, 1995.

[35] T. Ezzat and T. Poggio, "Facial analysis and synthesis using image-based models," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, 1996, pp. 116–121.

[36] Matrox, *Matrox Imaging Library, User Guide*, 1995.

[37] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, September 1990.

[38] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford Press, 1995.

[39] D. Machin, "Real-time facial motion analysis for virtual teleconferencing," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, 1996, pp. 340–344.